THE REAL RISKS OF ARTIFICIAL INTELLIGENCE

HOW TO HARMONIZE MACHINE & HUMAN VALUES



SHANNON VALLOR, SANTA CLARA UNIVERSITY

13TH CRO ASSEMBLY, ZURICH

SWISS RE CENTRE FOR GLOBAL DIALOGUE

NOVEMBER 27, 2017

Shannon Vallor

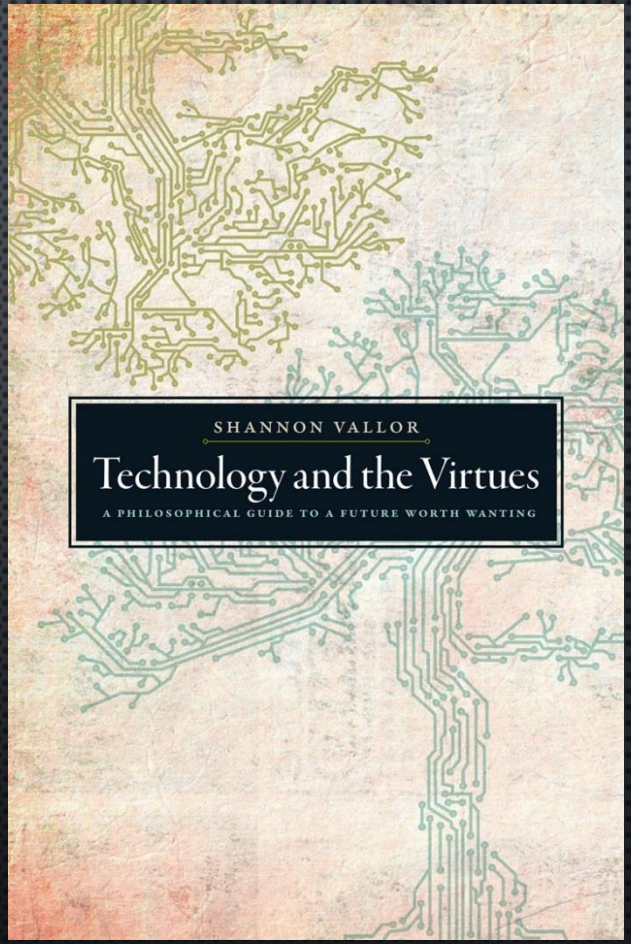
William J. Rewak, S.J. Professor
Department of Philosophy@SCU

Faculty Scholar
Markkula Center for Applied Ethics

Executive Board Member
Foundation for Responsible Robotics

Member, IEEE-SA's Global Initiative for Ethical Considerations in AI and Autonomous Systems

Technology and the Virtues: A Philosophical Guide to a Future Worth Wanting (OUP, 2016)





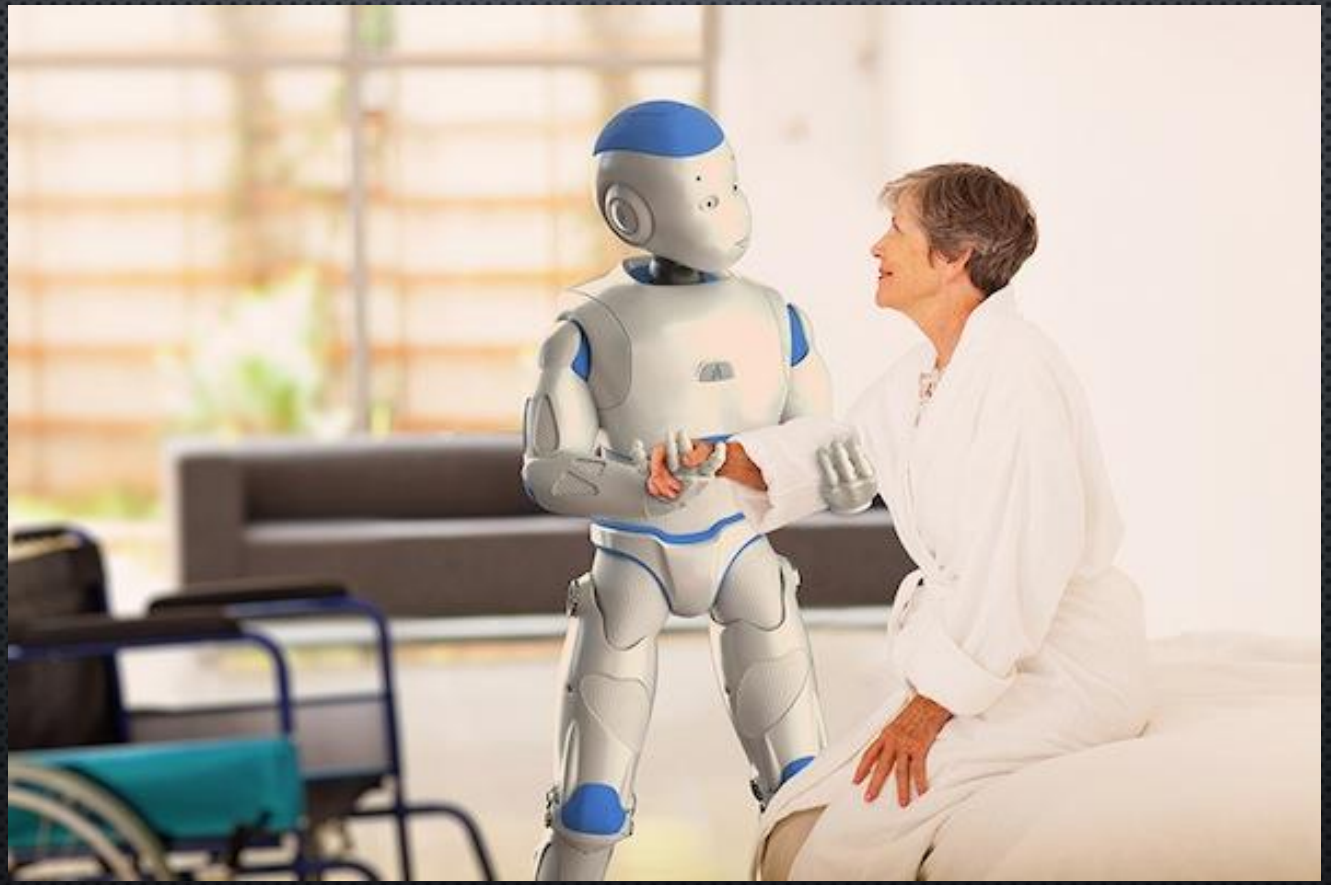
Self Driving Cars



VIRTUAL AGENTS/ASSISTANTS



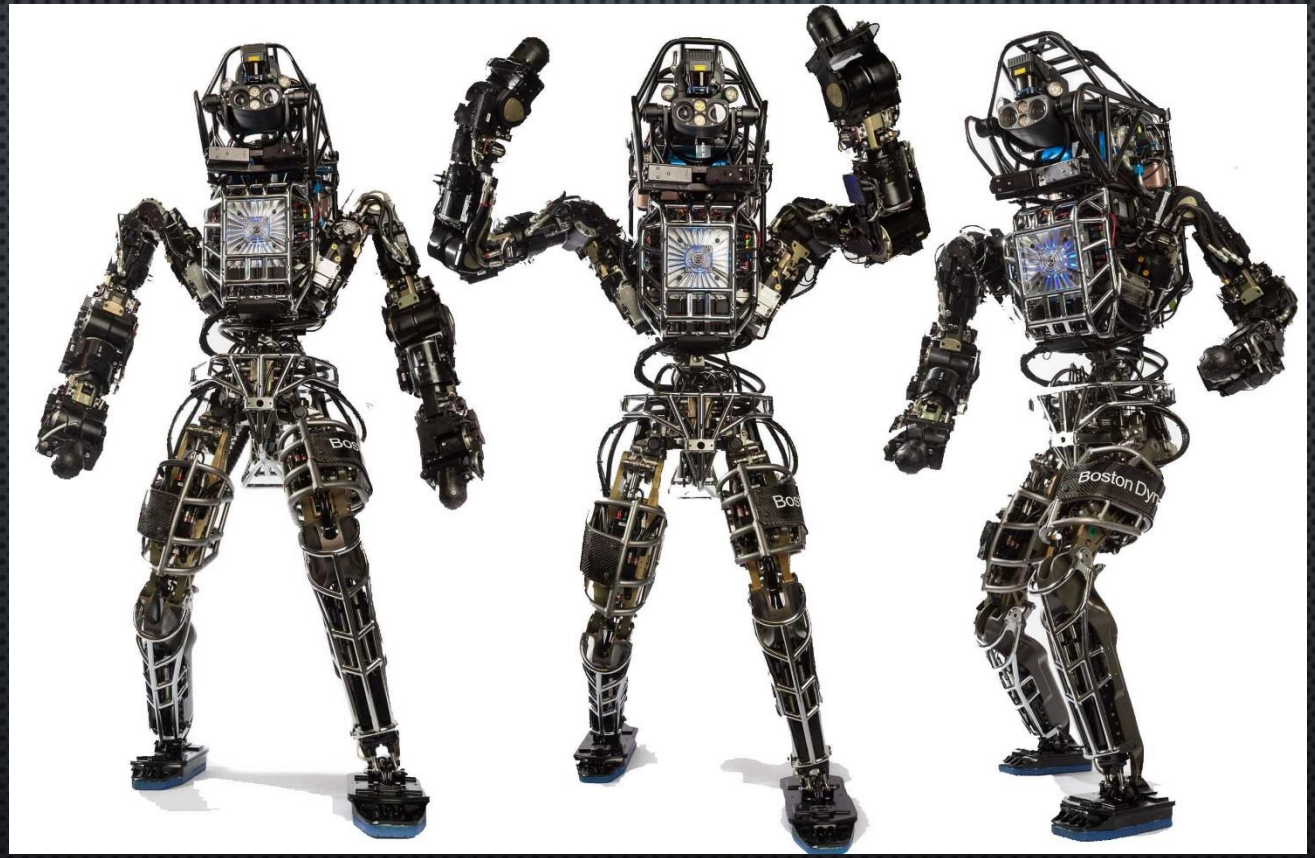
DECISION SUPPORT SYSTEMS



HEALTH CARE ROBOTS



Social Robots



MILITARY ROBOTS

AI AND RISK:

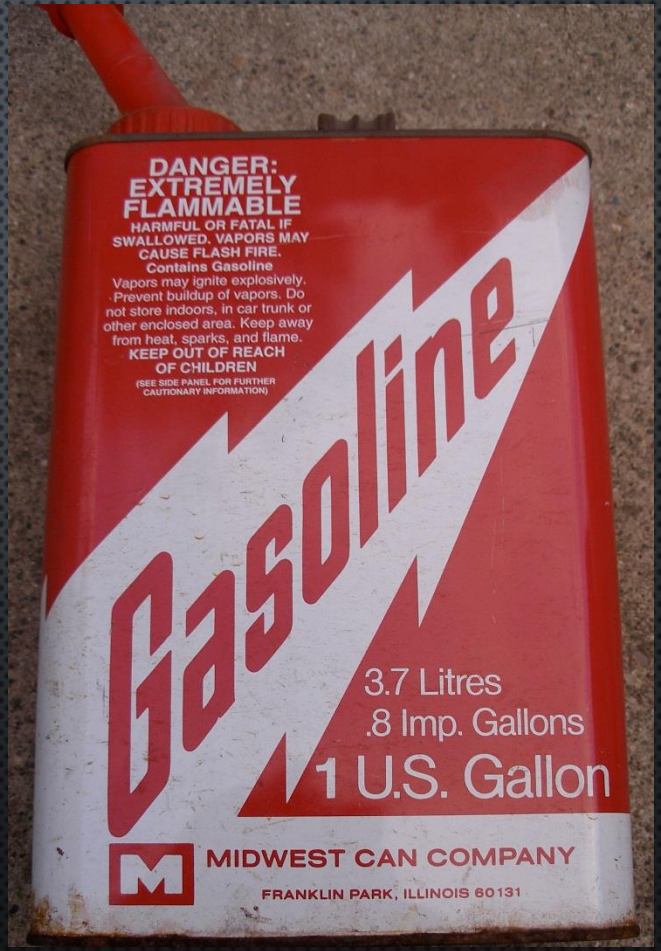
21st C. Artificial Intelligence will be a significant ***magnifier*** of economic, political, and physical risk.

21st C. Artificial Intelligence will be an important tool for ***managing & mitigating*** these same forms of risk.

Both of these things are true.

AI is a powerful ***accelerant –***

of beneficial and of destructive or destabilizing socioeconomic forces



AI AND RISK:

Where the *risk balance* ends up will depend not on how the ***technology*** develops, but on how quickly, broadly, and wisely ***human*** institutions and their leadership implement the reforms necessary to tip the balance in humanity's favor.

Defining A.I.

What Am I Talking About?

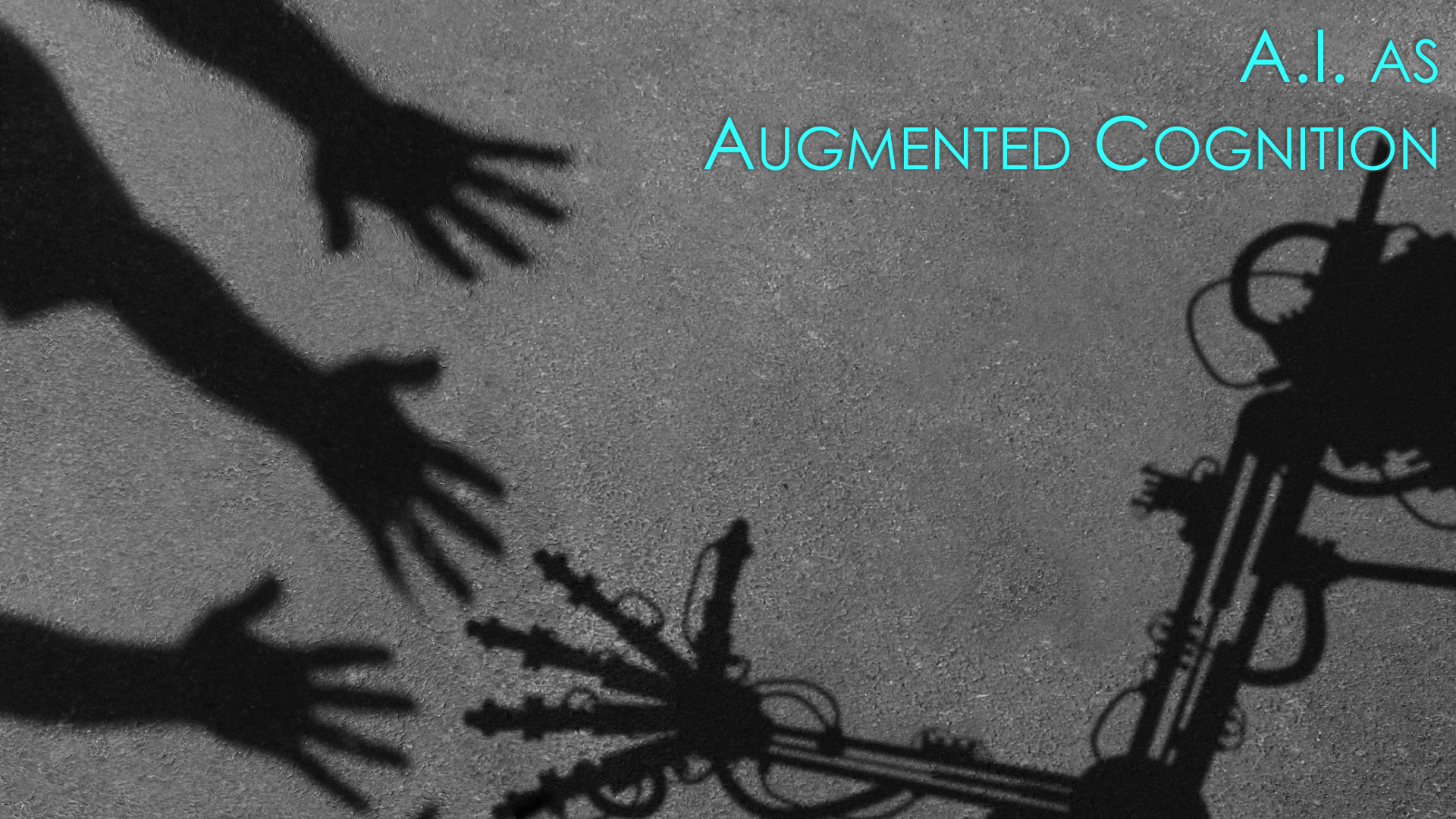
- Task-Specific A.I. (in contrast to **AGI**)
- Defined by Behavioral Competence **Only**
- Does **Not** Require System to Replicate Human Patterns of Thinking/Reasoning
- Frequently Entails **Machine Learning**

What Am I Talking About?

A.I. Does **Not** Require:

- **CONSCIOUSNESS** (SELF- AWARENESS)
- **UNDERSTANDING** (WORLDLY KNOWLEDGE)
- **SENTIENCE** (ABILITY TO FEEL/SUFFER)
- **JUDGMENT** (MORAL DISCERNMENT)
- **AGENCY** (RESPONSIBILITY, FREEDOM)

A.I. as Augmented Cognition





A.I. as Augmented Cognition

- A.I. *Can* Counter Harmful Biases in Human Thinking (**Only if Designed** With This Aim)
- A.I. Can Model Much Larger Space of Action Possibilities (**Surpasses Human Limitations** in Cognitive 'Bandwidth' and Speed)
- A.I. Not Bound by Human Cognitive Heuristics (**Discovers New Solutions** by New Methods)

Popular Visions of A Human-A.I. Future

FEARS

- Superintelligence
- Robot Overlords
- Skynet Scenarios
- A Jobless Future
- The WALL-E Dystopia

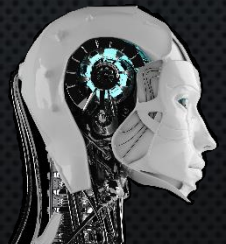


HOPES

- Humanity Amplified
- Humanity Unbound
- AI-Human Partnerships
- The Recovery of Leisure
- New ‘Age of Reason’

Does A.I. Pose *Existential* Risks to Humanity?

Yes – Just Not the Ones You've Heard About.



The Promise of A.I.

- New Scientific & Medical Discoveries
- More Efficient/Better Optimized Resource Use
- Reduced Human Drudgery / 'Busywork'
- Increased Business Productivity / Profitability



The Promise of A.I.

Renewal of Undervalued Human Skills:

- Empathy/Compassion for Others
- Social Acuity and Conflict Management
- Critical Thinking (Metacognition)
- 'Common Sense' (of Fairness, Humor, etc)
- Ethical Insight/Wisdom



Dimensions of A.I. Risk:

- Safety
- Transparency
- Privacy
- Autonomy
- Employment
- Accountability
- Power/Inequality
- Justice/Bias
- Diversity
- Human Vigor
- Wisdom



Key Questions in A.I. Ethics

How can A.I. be designed **safely** & **benevolently**, in ways that won't expose humans to unacceptable risks of harm, even when risky behavior may advance a system goal?

Key Questions in A.I. Ethics

How can we promote A.I. design that recognizes and gives appropriate weight to **human suffering**? How can we build machines that will refuse an illegal or unethical command? How can we build machines that humans can ***trust***?

Key Questions in A.I. Ethics

How can AI be designed to recognize and resist **hacking, abuse or malware** that aims to *weaponize* A.I. against vital human interests?

Key Questions in A.I. Ethics

How can we manage the risks to human well-being posed by **technological unemployment** from A.I. & robotic automation?

Key Questions in A.I. Ethics

How can we ensure that the economic benefits of A.I. are **justly distributed**, rather than amplifying the growing political and economic risks of extreme social inequality?

Key Questions in A.I. Ethics

How can we detect and mitigate the risks of **unjust discrimination** & **harmful bias** in machine judgments, which may already be skewing decisions in hiring, lending, & jurisprudence?

Key Questions in A.I. Ethics

Should any artificial agent occupy a '**position of responsibility**?' What would that mean? Can machines be held *responsible* for actions they take without human direction? Can a machine ever *understand and accept* ***risk?***

Key Questions in A.I. Ethics

Is **meaningful human control** of AI possible?
Who should **steer** or **govern** A.I. technology?
How can we promote the ***responsible* use of A.I.?**

Key Questions in A.I. Ethics

What kind of **social roles** should artificial agents occupy?

Caregivers for the elderly/sick? For our children?
Teachers? **Priests**? **Therapists**?
Law Enforcement Agents? **Soldiers**?
Legal/Judicial Advisors?
Romantic partners? **Friends**?

What are the **risks** of allowing artificial agents to occupy these various roles? Do the advantages **justify** the risks?

What cognitive labor *WON'T* A.I. be able to do for us, work that we will need *more* humans to do?

Answer questions like these.

If we do it *right*, life *with* A.I. can be more *humane* than life without it

Instead of repetition, calculation, statistical analysis, and prediction (things humans are only *modestly* good at)

humans will employ creative, social, ethical, & critical thinking skills to identify and bridge the gaps between machine intelligence and ours

wisely selecting the greatest benefits of A.I. and minimizing/mitigating the risks

BUT…

This future of human-A.I. partnership, one that serves and enriches human lives, won't happen organically;

it will need to be a *choice* we make.

Will companies and governments *invest* adequately in the human capital needed to fill these roles?

Or will short-sighted profit-seeking in the first waves of A.I. deployment lead to:

*under*investment in human guidance,

degraded A.I. safety & performance,

& damaged public trust in A.I./tech?

Will corporations, lawmakers, & schools invest enough in new skills training for these roles?

Will we be flexible enough to adapt our educational & economic habits to this need?

Will we create *enough* opportunities for humans to find meaningful and rewarding work in an A.I.-driven economy?

Will we rise to the occasion and seize the opportunity that A.I. presents?

Or will institutional & social inertia lead us to make the same short-sighted mistakes we made with other modern technologies (coal, pesticides, plastics, social media, etc.)?


STEEP
GRADES
SHARP
CURVES
NEXT 4 MILES
25
M.P.H.

What kind of future can we expect from A.I.?

That's still very much up to us.

What kind of future with A.I. should we *want*?

One in which *human* values remain central.

But *which* human values?

Understanding A.I. Risk: Human & Machine Values



What is a ‘Machine Value’?

A value driven by the cultural context of mechanized industry and science

A value that is easily represented algorithmically & operationalized in computational systems

A value indicating an economic or cognitive advantage of machines over humans

A ***misnomer*** **–** ‘machine values’ *are* human values, just ones that suit themselves well to implementation in machine cognition.

A.I. is us, all the way down.

Machine Values

- Optimality
- Efficiency
- Speed
- Precision
- Classification
- Identification
- Discrimination
- Readability
- Compressibility
- Replicability
- Invulnerability
- Reliability

When Machine Values Attack

The New A.I. Phrenology

Facial or other biometrics are used for social surveillance and as criminal/risk indicators – often in reckless ways that generate spurious correlations and false positives that can ruin lives

Bias-Laundering

A.I. training data rife with racial, gender, class biases are being misused to provide *seemingly* 'objective' DSS for judges, employers, bankers, teachers, etc.

Gaming Vital Human Systems

AI is used as ‘arbitrage’: ruthlessly exploiting machine advantages over humans in order to manipulate, undermine, or wrongfully profit from vital human institutions (media, voting, finance, etc.)

‘Black Swan’ Events

As complex adaptive systems, even statistically reliable ML systems used ‘in the wild’ will generate rare but wildly unexpected ‘emergent behaviors’ that we do not want and are wholly unprepared for

Are we as prepared to manage these risks of A.I. as we are to reap its rewards?

CASE STUDY:

TECH

AI experts slam proposed use of ‘extreme vetting’ algorithms

WRITTEN BY
Tajha Chappellet-Lanier

NOV 17, 2017 | FEDSCOOP

WRITTEN BY

Tajha Chappellet-Lanier

NOV 17, 2017 | FEDSCOOP

While the extreme vetting initiative has not yet begun, U.S. Immigration and Customs Enforcement is actively looking for a contractor to deliver a service that "automates, centralizes and streamlines the current manual vetting process." According to the Brennan Center for Justice, ICE wants to award a contract by September 2018. The directive comes from President Donald Trump's January 2017 executive order which calls for a screening process to "evaluate the applicant's likelihood of becoming a positively contributing member of society and the applicant's ability to make contributions to the national interest."

Herein lies the problem, experts say — "As far as we are aware, neither the federal government nor anyone else has defined, much less attempted to quantify, these characteristics," they write. "Algorithms designed to predict these undefined qualities could be used to arbitrarily flag groups of immigrants under a veneer of objectivity."

Which machine values are driving this call?

Classification, Prediction, Discrimination

What human values have been jeopardized, demoted, disregarded?

Justice, Compassion, Wisdom

What kind of future might this use of A.I. tend to enable? Is it one we should want? Is it one we should be working *toward*?

Looking in The A.I. Mirror



What Makes an Artificial Intelligence Rac
Sexist

Google Creates Chatbot That Thinks It's Immoral To Have A Child

Robots 'Evolve' the Ability to Deceive

rs discover why AI
exist

AI systems have a self-taught sexist view of women

What did women do to deserve all this?

teaching bots art of
the AI learned to lie

Unethical AI

ARTIFICIAL INTELLIGENCE CAN BE SEXIST AN
RACIST JUST LIKE HUMANS

BY ANTHONY CUTHBERTSON ON 4/14/17 AT 12:22 PM

Live Science › Tech

Bad News: Artificial Intelligence Is Racist, Too

Racist & sexist AI bots could deny you job, in
tech experts

Rise of the racist robots - how AI is learning all our worst impulses

REVEALED: AI is turnir
RACIST as it learns from
humans

NEWS

Even Artificial Intelligence Is Sexist

Biased Algorithms Are Everywhere, and No One Seems to

THE HARD TRUTH

AIs are incapable of fear, hate, the desire to dominate and oppress, or any of the destructive emotions of humans.

They can only reflect our own.

But they do so *very, very* well.

The A.I. Mirror

- All designed human artifacts are **mirrors of human value –** artificial agents included.
- Just as the image in my mirror is still ***my own***, our A.I. mirrors are still ***us***, only **externalized** and **transformed**
- A.I. mirrors may **magnify**, **reduce**, **invert**, **distort**, **center**, or **decenter** human values

RESPONSIBLE A.I. DESIGN AND USE takes deliberate account of these mirror effects and their ethical impact on human lives and institutions.

Question: if you look in a mirror and see a dirty face looking back, what should you do?

A. Break the mirror

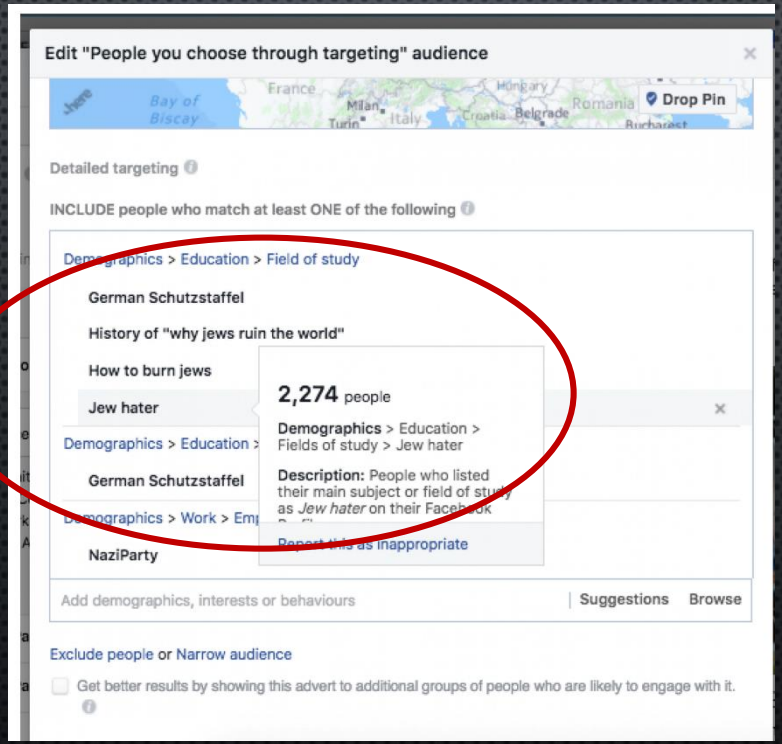
B. Start an anti-mirror campaign

C. Start scrubbing the mirror

D. Accept what the mirror shows as 'inevitable'

E. Try washing some dirt off your face

THE A.I. MIRROR IN ACTION



Edit "People you choose through targeting" audience



Detailed targeting

INCLUDE people who match at least ONE of the following

Demographics > Education > Field of study

German Schutzstaffel

History of "why jews ruin the world"

How to burn jews

Jew hater

2,274 people

Demographics > Education > Fields of study > Jew hater

Description: People who listed their main subject or field of study as *Jew hater* on their Facebook

Report this as inappropriate

Demographics > Education >

German Schutzstaffel

Demographics > Work > Emp

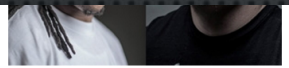
NaziParty

Add demographics, interests or behaviours | Suggestions Browse

Exclude people or Narrow audience

Get better results by showing this advert to additional groups of people who are likely to engage with it.

THE A.I. MIRROR IN ACTION

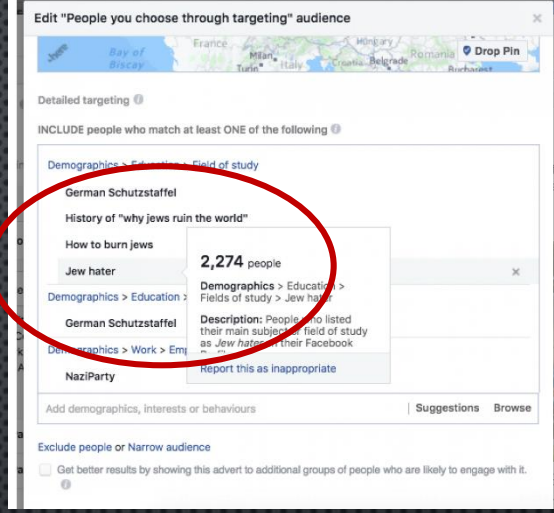


MACHINE BIAS
Investigating
Algorithmic Injustice

Until this week, when we asked Facebook about it, the world's largest social network enabled advertisers to direct their pitches to the news feeds of almost 2,300 people who expressed interest in the topics of "Jew hater," "How to burn jews," or, "History of 'why jews ruin the world.'"

To test if these ad categories were real, we paid $30 to target those groups with three "promoted posts" — in which a ProPublica article or post was displayed in their news feeds. Facebook approved all three ads within 15 minutes.

After we contacted Facebook, it removed the anti-Semitic categories — which were created by an algorithm rather than by people — and said it would explore ways to fix the problem, such as limiting the number of categories available or scrutinizing them before they are displayed to buyers.



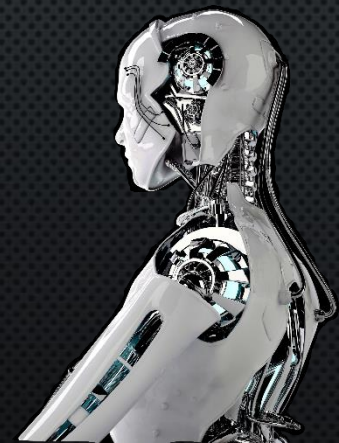
IS THIS CLAIM ENTIRELY TRUE?

In what sense **isn't** it true?

What ***harm*** is there in saying:

'these categories were created by an algorithm rather than by people'?

Managing A.I. Risk



Invest in Human-AI Partnerships: Human wisdom is the key asset for *baking human values* into AI/ML design and development, and ensuring their promotion in AI implementation and use

Plan for Failure: Even under ideal conditions, the social implementation of AI cannot be *risk-free*. This will *not* be a smooth ride. We need to anticipate rare but dangerous AI-driven failures just as we anticipate and build in safety margins for brake and engine failures, for blown tires and overheating reactors.

ASK THE RIGHT QUESTIONS:

- **Which human values** will this AI system reflect, magnify, reduce, or distort?
- **Which** human values could it eclipse or degrade, especially downstream or in the wrong hands?
- **Who** may be harmed if that happens? **How?**
- **What** are our harm mitigation and human value preservation strategies?

DESIGN AI FOR HUMAN VALUES:

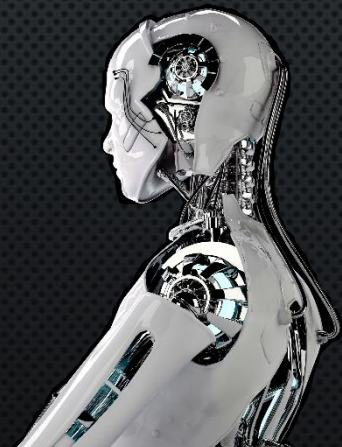
- Beware of '**Reverse Adaptation**' (Sullins 2014)
- **Audit A.I. outcomes** using a broad value spectrum – aim to ***conserve*** value complexity, not reduce it
- Find **design workarounds/allowances** for human values not well expressed in algorithms: Compassion, Justice, Fairness, Love, Hope, Responsibility, Liberty, Dignity, etc.
- **Institutionalize** Ethical A.I. Design & Use

FOCUS ON THE REAL RISKS:

Existential risks to humanity include more than just threats to our species' survival; they include any risk to a future for our species that is ***worth wanting***.

There ***aren't*** any plausible scenarios for this century where hostile A.I.s willfully destroy us. There ***are***, however, many plausible scenarios where ***human*** misuse of A.I. destroys our best chances of a future worth wanting. ***These*** are the risks we need to invest most heavily in avoiding.

Lessons from the A.I. Mirror



1. The ***real*** risks of A.I. *aren't* those of 'superintelligent' machines taking over.

The real risks will come from machines that aren't yet ***smart enough*** to handle the responsibilities humans will give them.

2. Machine intelligence will keep improving but will remain far ***narrower*** than ours for the foreseeable future;

This will boost demand for human skills and values that can ***bridge the gap*** and mitigate the ***risks*** of narrow but powerful A.I. acting on broad, complex human societies

3. Real machine ***minds*** aren't coming any time soon. But machines that ***act*** like minds (up to a point), are already here, and these fictions will only become more convincing.

This poses **immense risks** to humans.

4. Investment in human ***enrichment*** and ***responsible oversight*** of AI are the keys to managing these risks, & sustaining growth and public trust in the technology.

AI is ***not*** ready for solo flight.

5. Machine values too often eclipse broader human values. To ***harmonize*** them, we must take ownership of both, making the former serve the latter—

never *the reverse*.

6. AI must not become a scapegoat for **human** moral failures, even those that it may reflect, magnify, & perpetuate.

We are still the responsible agents.
This is ***good*** news.

7. The AI mirror is a powerful **diagnostic** tool for society – we can learn as much from its failures and limitations as we can from the new insights and solutions it delivers.

AI can help keep us ***honest*** about who we are and what we choose to value.

8. AI can be a powerful force for good and a driver of human growth; trying to make ***it*** better can also make ***us*** better.

What good can we do ***with*** AI that we *couldn't* do ***without*** it?

9. Nothing is inevitable unless we make it so.

The future of AI, and of human society, will not be determined for us, but ***by*** us.

Elon Musk and Steven Hawking got it wrong.

A.I./Robot Overlords **aren't** coming to choose our future for us. It's still on us to get it right, and to decide what 'right' looks like

It’s always going to be ***us*** in the AI mirror.

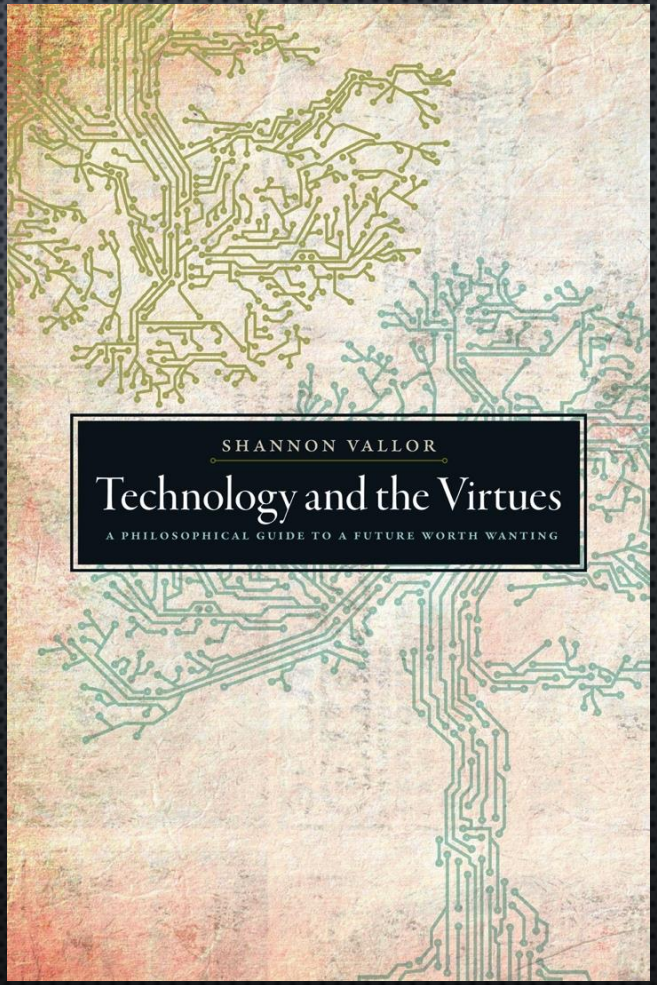
And it’s still ***up to us*** whether we’ll like what we see.

THANK YOU!

@ShannonVallor

svallor@scu.edu

www.shannonvallor.net



Basic Copyright Notice & Disclaimer

©2017 This presentation is copyright protected. All rights reserved. You may download or print out a hard copy for your private or internal use. You are not permitted to create any modifications or derivatives of this presentation without the prior written permission of the copyright owner.

This presentation is for information purposes only and contains non-binding indications. Any opinions or views expressed are of the author and do not necessarily represent those of Swiss Re. Swiss Re makes no warranties or representations as to the accuracy, comprehensiveness, timeliness or suitability of this presentation for a particular purpose. Anyone shall at its own risk interpret and employ this presentation without relying on it in isolation. In no event will Swiss Re be liable for any loss or damages of any kind, including any direct, indirect or consequential damages, arising out of or in connection with the use of this presentation.